PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 1003–1014
doi:10.1098/rsta.2008.0237
Published online 16 December 2008

Information in environmental data grids

By B. N. Lawrence[1,*], R. Lowry[2], P. Miller[3], H. Snaith[4] and A. Woolf[1]

[1]*Rutherford Appleton Laboratory, STFC, Didcot OX11 0QX, UK*
[2]*Proudman Oceanographic Laboratory, British Oceanographic Data Centre, Liverpool L3 5DA, UK*
[3]*Plymouth Marine Laboratory, Plymouth PL1 3DH, UK*
[4]*National Oceanography Centre, Southampton SO14 3ZH, UK*

Providing homogeneous access ('services') to heterogeneous environmental data distributed across heterogeneous computing systems on a wide area network requires a robust information paradigm that can mediate between differing storage and information formats. While there are a number of ISO standards that provide some guidance on how to do this, the information landscape within domains is not well described. In this paper, we present an information taxonomy and two information components, which have been built for a specific application. These two components, one to aid data understanding and the other to aid data manipulation, are both deployed in the UK NERC DataGrid as described elsewhere.

Keywords: geospatial data grid; metadata; Open Geospatial Consortium

1. Introduction

Data grids can be defined as distributed computing systems that provide access to data distributed across the nodes of the computing system. There are a multitude of instances of data grids, which differ across all the possible dimensions inherent in the definition: different ways of distributing the data; different types of data; and different methods of access. One classification of such data grids is to consider the difference between homogeneous data and services that are distributed for efficiency or economic reasons, and pre-existing heterogeneous data being aggregated via middleware, which provides homogeneous interface(s). In this paper, we describe the information requirements for one specific instance of the latter, an environmental data grid constructed for the UK Natural Environment Research Council (NERC)—the NERC DataGrid (NDG). Although grounded in one application, the work described here should have general applicability.

(*a*) *Background*

Much previous work on data grids and data interoperability has been predicated on building solutions where homogeneous software for storage (e.g. SRB; Wan *et al.* 2003) or middleware (e.g. Globus; Foster *et al.* 2006) can be deployed. The former

* Author for correspondence (bryan.lawrence@stfc.ac.uk).

One contribution of 24 to a Discussion Meeting Issue 'The environmental eScience revolution'.

was not relevant to our situation with pre-existing data management solutions, and the latter did not address the key problem facing the NERC community: semantic interoperability. In our context, key attributes of the data are the syntax (format and layout), the semantics (meaning) and the 'grid solution' needed to address them. Such attributes may be inherent in data, or appear in metadata: for example, the syntax of a binary formatted file may only be understood by using a metadata document (the original program that wrote the data), or it may be inherent in the data because it conforms to a well-known format (e.g. a .png image format or a .nc NETCDF[1] file). Similarly, whether the data are a vertical profile of temperature or a four-dimensional gridded humidity mixing ratio, such semantics may be documented within the file (implicit in a CF[2] compliant NETCDF file) or externally documented in a text file. Obviously both can be true; inherent metadata can be replicated in external metadata.

The general scope of the problem is well described in ISO 19101 (2002), which describes and categorizes both the interoperability issues and the nature of (some of) the information entities. While ISO 19101 (2002) was not written from a grid perspective, the underlying concepts are identical: ISO 19101 (2002) talks about a set of layers building up from network protocol interoperability, through file system interoperability, to syntactically aware applications and then semantically aware applications.

ISO 19101 (2002) also outlines the key information entities relevant to services aimed at exposing data. Those key information entities are characterized in figure 1: the key points are that a dataset can itself contain multiple datasets, each of which can be thought of as an aggregation of 'features' each of which may have some characterization and spatio-temporal location. Features themselves are named entities for which 'coverages' (fields of numbers over some spatio-temporal domain) may exist. External metadata may exist for a dataset, and services that operate on the data may exploit the external metadata, and are themselves described by metadata. The network address of the data and the services is a key property to be understood, and may differ between dataset(s) and service(s). Although we do not discuss them further here, key properties of the data are also those needed to support specific services.

(*b*) *Constraints*

In choosing solutions for environmental grids, which are exposing heterogeneous pre-existing data to different communities, there are two strong constraints: (i) data holders cannot, and will not, consider changing the way they store data and fundamental metadata (including user information) and (ii) data users will not have the time and inclination to exploit foreign (to their discipline) formats and information infrastructures. These constraints lead to the conclusion that interdisciplinary exploitation of data is predicated on both interdisciplinary exploitation of the information about data and hiding the complexities of native data formats. Achieving both of these requires new 'information' middleware.

[1] The network Common Data Form, see http://www.unidata.ucar.edu/software/netcdf/.
[2] The climate forecast conventions for NETCDF are documented at http://www.cfconventions.org.

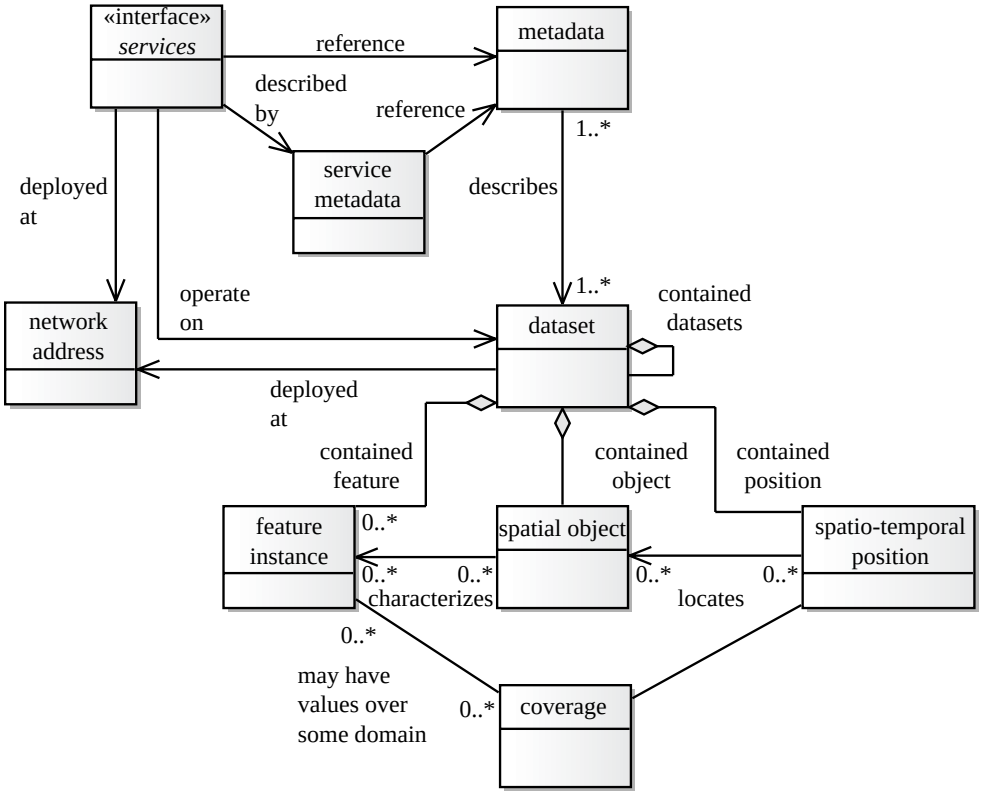

Figure 1. Key information entities (extended from ISO 19101 (2002)). (All figures, except figure 2, are class diagrams constructed in the Unified Modelling Language, UML.)

2. The information environment

Understanding the information environment is a key to building common access to heterogeneous data. ISO 19109 (2005) suggests a methodology based on building a model of the information, implementing an XML schema description, and requiring that any given information document should be an XML instance of that schema. The NDG experience was that such an approach applied to the entire information spectrum required for an environmental data grid would have been enormously taxing and would not have exploited the knowledge and experience available in the existing metadata structures. However, using this approach with two key modifications was practicable: if the metadata spectrum was broken into a taxonomy, tailored solutions for each class of metadata could be constructed, and if the XML instances could integrate with the existing data formats and encode only the key semantic structures necessary for interoperability, then the size of the task was practicable.

The NDG metadata taxonomy (figure 2) consists of the following key classes of metadata.

— *A-archive* metadata describes the syntax and semantics (e.g. parameter descriptions) of the data objects themselves. The concept is further described below in §2*b*.

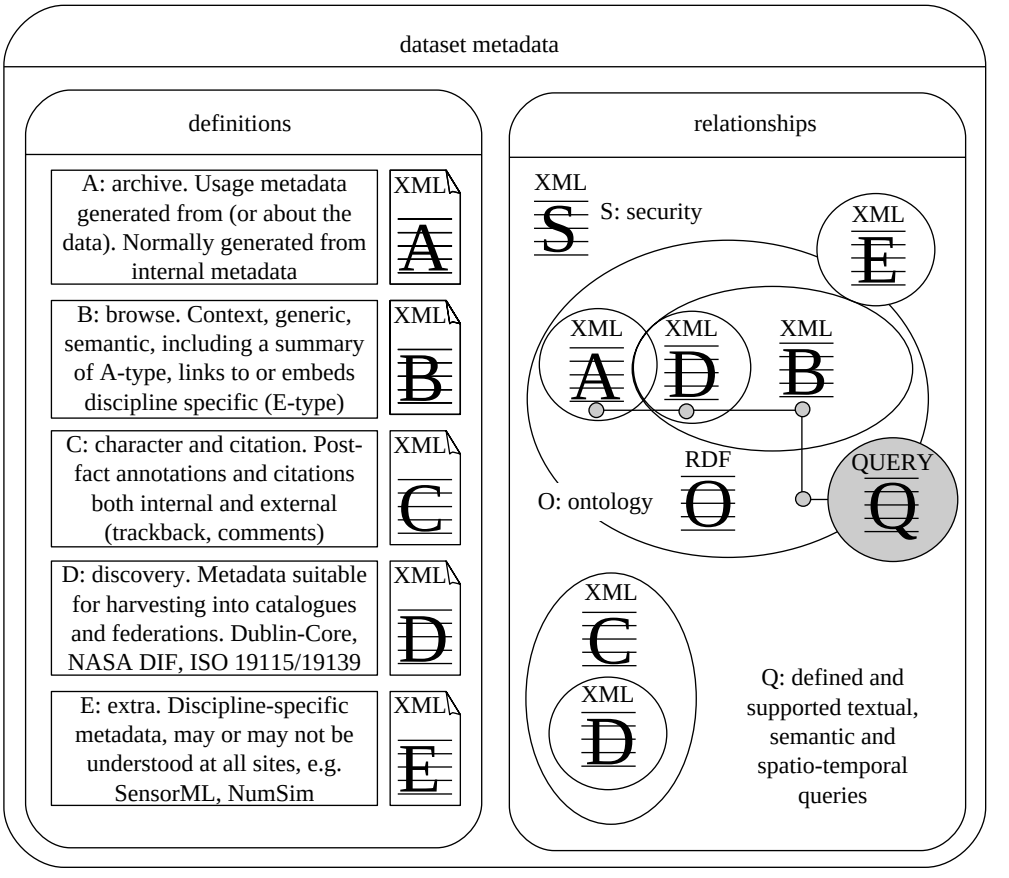

Figure 2. Dataset metadata taxonomy.

— *B-browse* metadata supports understanding the context of data and choosing between similar datasets. This concept is further described below in §2*a*.
— *C-character* metadata includes citations of the data itself, and post-fact assertions as to the quality of the data. Typically, such metadata does not always exist packaged with the data itself, but may exist in third party repositories (e.g. journal archives), etc. (Note that C-metadata itself may be discoverable by D-metadata.)
— *D-discovery* metadata is a subset of the browse and archive metadata, which is selected to aid finding data for evaluation or visualization and/or other uses. Typically discovery metadata is harvested and/or submitted to other organizations to aid data discovery.
— *E-extra* metadata is the core discipline- or instrument-specific metadata, which may be strongly typed (i.e. conforms to schema such as SensorML[3]) or consist of arbitrary documents. Providing consistent interfaces from B-metadata to E-metadata was one of the main challenges identified for the NDG).

[3] The Sensor Model Language, see http://www.opengeospatial.org/standards/sensorml.

Other metadata components included a register of the relationships between data entities and their access control (S), the descriptions of the supported query types (Q) and ontologies, which link semantic components together (O). The latter is a key to interdisciplinarity: one group's 'rainfall' is another group's 'precipitation'.

In the remainder of this section, we present the information schema used within NDG to support browse and archive metadata as both were constructed specifically for use in describing environmental data for data grids.

(*a*) *Browse: metadata objects for linking environmental sciences, MOLES*

The key step between finding and using data is understanding, putting in context and potentially choosing between very similar data offerings, hence the term B-browse in the metadata taxonomy. Often the same information that is used for context and understanding can form the basis of the discovery. For example, one might arrive at a dataset via a search for data produced by a specific tool—and then need to understand the context of where that tool was deployed. Similarly, one might find the very same data from the perspective of the observing location, and then need information about the tool. In both cases, the information requirements for context and understanding may well exceed those of discovery. These relationships and the underlying information essentially flesh out the detail required to implement the metadata component of figure 1, and are encoded in the metadata objects for linking environmental sciences (MOLES) data model as shown in figure 3.

The MOLES supports a number of first-class entities, which together provide linkage between key characteristics of the description of data. The key entities are the dgProductionTool, which characterizes the instruments and/or processes available for producing data; dgObservationStation, which characterizes the location(s) (and observers) of data production; dgActivity, which characterizes the projects and campaigns etc., associated with data production; and dgData itself, which consists of aggregations of more dgData entities, or of dgGranules, which are each associated with one instance of A-type metadata for a dataset (note that this is an OR, not an AND; data entities cannot consist of other data entities and granules!). These four key entities are related by the dgDeployment, which binds a production tool deployed at an observation station on behalf of an activity to produce data. Additionally dgService entities can manipulate the other high-level entities to produce either new high-level entities, or HTML output (text and/or visualizations). All high-level entities can be described by D-discovery metadata (currently we have implemented the NASA GCMD DIF,[4] but future versions will use ISO compliant metadata (ISO 19115 2005)), although in practice, experience suggests that exposing data granules individually actually worsens user experience of data discovery both in speed and relevance—users navigate to the data they want far more quickly by navigating to a data entity (e.g. a model simulation) and then to a data granule (e.g. the monthly mean output as opposed to the daily mean output), rather than trying to choose between many (hundreds to thousands) of data granules that may only differ in small ways. Additional relationships (O-type metadata) between the high-level entities can also be recorded via the relatedTo association class. Currently, these relationships are not used, but the next version of MOLES

[4] NASA Directory Interchange Format (DIF), see http://gcmd.nasa.gov/user/difguide/.

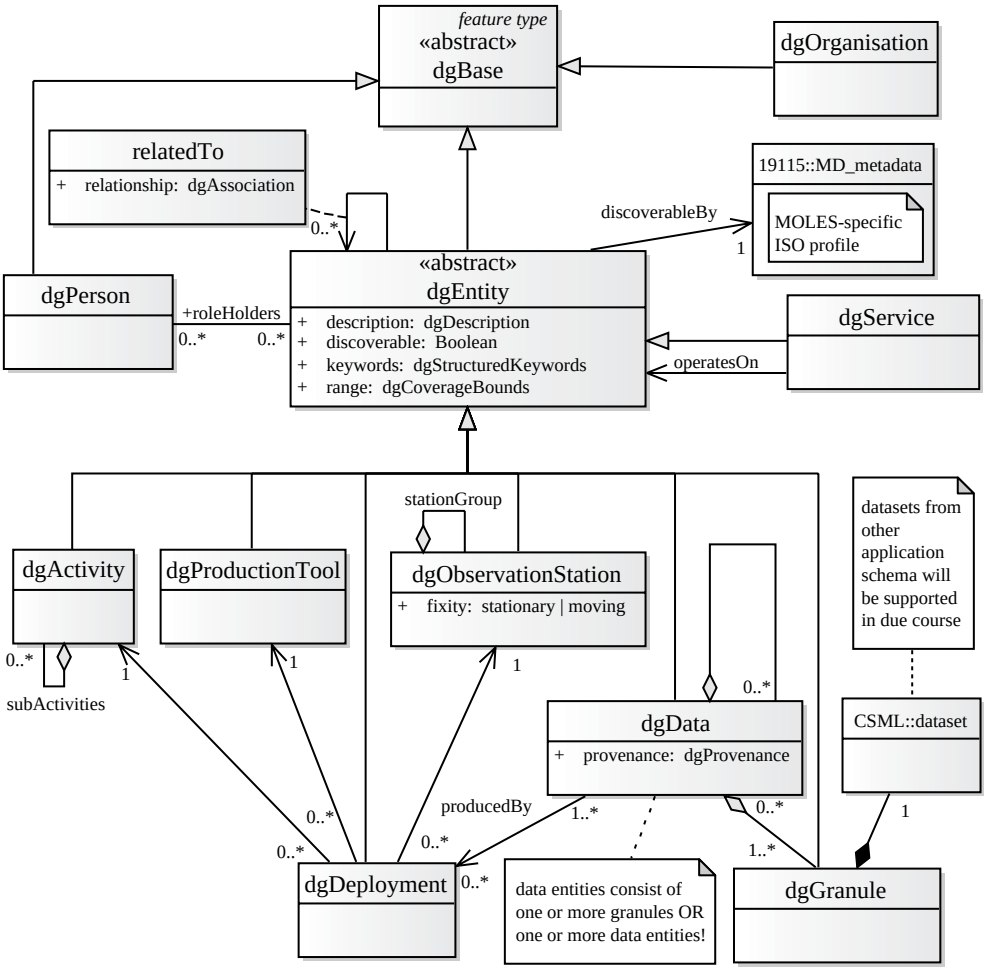

Figure 3. Basic concepts of MOLES. Compare with the ISO framework in figure 1 and the taxonomy in figure 2. Only important fundamental class attributes are shown.

will implement Resource Description Framework technologies to use them to support navigation.

Apart from the data granules, which are strongly typed by the use of the application schema, and the service descriptions for which strong typing (and thus service orchestration) is desirable (but not yet implemented), the other high-level entities are relatively weakly typed. General attributes consist of relationships, some limited code lists to aid ontology-based navigation, and typed links to documents such as (in the case of productionTools) SensorML documents. Typed links from B-metadata to E-metadata documents are desirable because presentation services can then be developed, which can use subsets of the target documents and/or produce navigation aids to assist in finding key information. One important attribute is provenance information which we here define as inclusive of both automatic and human-generated process metadata: individual process step descriptions could appear with the dgProvenance attribute of data entities (as shown here) and internal to the process metadata within dgProductionTool.

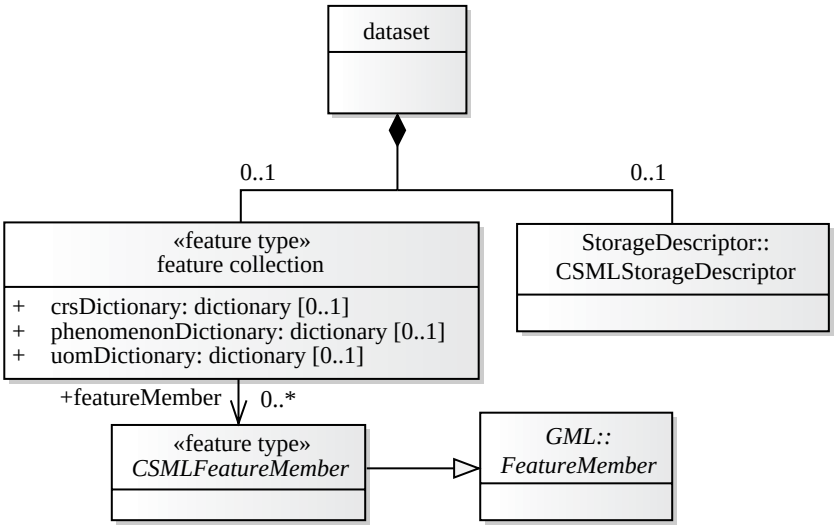

Figure 4. CSML dataset.

(*b*) *Archive: the climate sciences modelling language*

The key requirements of A-metadata are that they provide a complete description of the data, constructed in a manner which can, where appropriate, be used by different communities to obtain the data in their own native formats. The key semantics are those described in figure 1: the location of the measurement (or simulation), the details of the feature itself (including a coverage description where appropriate), and the syntactical layout of the data itself. Key external dictionaries are enumerated as attributes of the feature collection.

As described above, the ISO methodology is to exploit data modelling to construct an application schema of the Geographic Markup Language, GML (ISO 19136 2005). An application schema is essentially a profile of GML which restricts and extends GML in a manner appropriate for describing a range of features of interest to a particular community. The initial focus of the NDG was to support interoperability within and between the oceanographic, meteorological and remote sensing communities, and so NDG constructed a specific application schema targeted to that usage—the climate science modelling language, CSML. By exploiting this methodology, semantics can be shared with other communities: for example, the international geology community (including the British Geological Survey) have also constructed an application schema of GML, so in principle it ought to be possible to extract and exploit together appropriate semantics from instances of either schema.

In building CSML, it was necessary to recognize two specific limitations of the existing GML methodologies: scalability and coordinate support. With terabyte scale pre-existing archives, it is not possible to encode the data itself in XML documents and GML has poor support for coverage types with irregular geometries, complex time domains or unusual vertical coordinates. CSML (figure 4) recognizes these two limitations by (i) including within a CSML dataset an XML storage descriptor document that provides a methodology for linking out

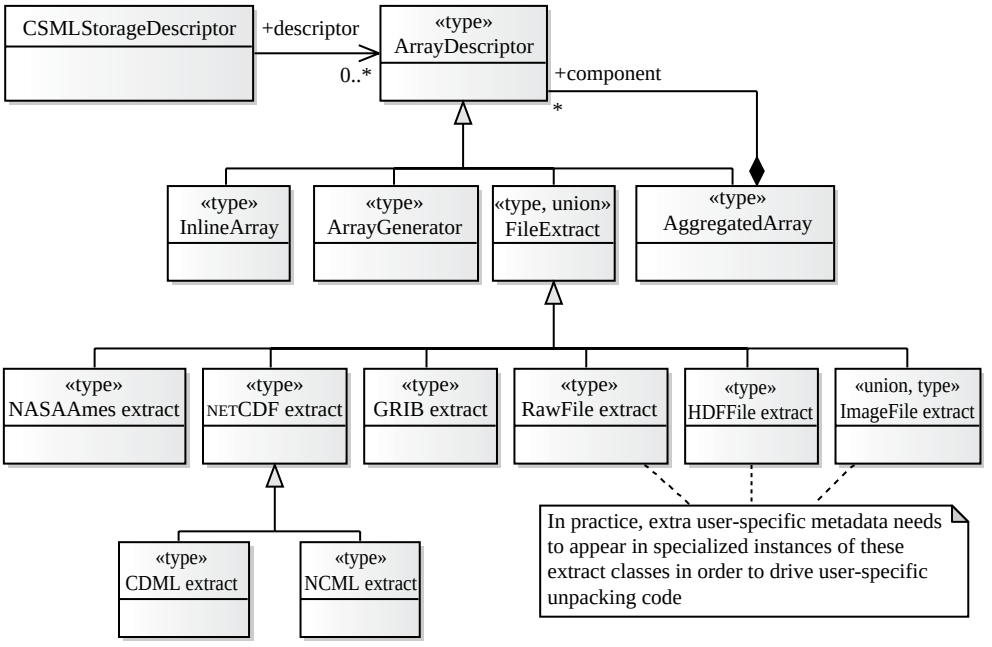

Figure 5. CSML storage descriptor.

from the CSML features (which themselves specialize GML features) into binary storage (which might exist either because of volume issues or simply because of legacy), and (ii) implementing a more complete (ISO 19123 2005) coverage specification (rather than just the limited form currently included within GML).

The key problem to be resolved in the storage descriptor (figure 5) is to efficiently describe arrays so that services can be used to exploit the data (and potentially reformat for reuse in different communities). The storage descriptor does this by supporting a range of solutions: from inlining the actual XML (as is generally done with most applications of GML), to algorithmic support for generating an array (as one might do for an equally spaced coordinate array), to providing a description of external storage, or aggregations of the previous. Where the data exist in external storage, there are three major categories: (i) storage in external files that have strong enough conventions that library code can extract (and write data), examples include NASA Ames,[5] NETCDF (with CF conventions) and GRIB,[6] (ii) external storage in files that would require bespoke semantic descriptions (and code) to extract the arrays of interest, examples include HDF[7] without conventions, generic image files and arbitrary raw binary files, and (iii) external storage in databases (not shown), but supportable through a bespoke SQL schema aware specialization of RawFile extract. (In practice, grid middleware such as the GLOBUS data access and integration layer (OGSA-DAI;

[5] NDG currently supports v. 1.3 of the Gaines-Hipskind format, see http://espoarchive.nasa.gov/archive/docs/formatspec.txt.

[6] The World Meteorological Organisation's GRidded Binary format, see http://www.wmo.ch/pages/prog/www/WDM/Guides/Guide-binary-2.html.

[7] The Hierarchical Data Format, see http://hdf.ncsca.uiuc.edu/.

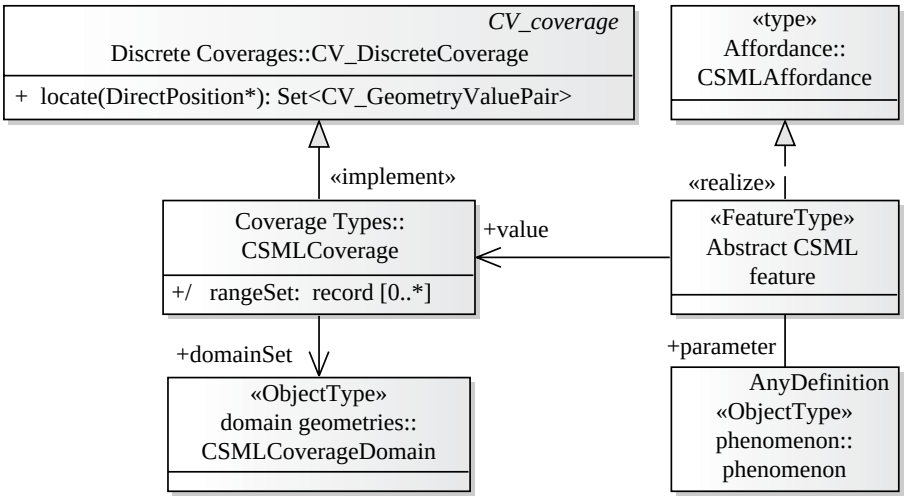

Figure 6. CSML abstract feature. All CSML features follow this pattern. (Note that functional methods are discriminated from attribute values by names, which include brackets.)

e.g. Karasawas *et al.* 2005) could also be deployed as specializations of raw file extract, but they too would need a semantic description layer to fit into the storage descriptor.)

While GML implements feature types, they are implemented entirely as objects with attributes but no methods: the design of CSML recognizes this by extending feature types so that they realize types that afford methods (they have 'affordances'; figure 6). Apart from these exceptions, the figures presented here showing CSML structure conform to the specifications of ISO 19103 (2005) as implemented in GML.

CSML (v. 2) implements 13 specific feature types: point data; time series of point data; trajectory data; point collections; profiles; profile series; ragged profile series; section; ragged section; scanning radar; grid; gridseries; and swath.[8] This set of feature types was selected as a trade-off between excessive specificity (for example, a specific feature type for a radiosonde which is different from a dropsonde) and too much generality (CSML v. 1 had only seven feature types and using these types required too many special case adjustments). The way these map onto the abstract feature type can be seen by comparing figures 6 and 7. The latter shows the characteristics of the ragged section type, which was designed to support the sort of data obtained by ships dropping multiple casts of instrument(s) to varying depths during a cruise. The nature of the functional affordances is clear: these are the meaningful subselections one can carry out on a ragged section, and they support the subselections offered by the services described in Latham *et al.* (2009).

A key part of designing for interoperability is recognizing appropriate levels of governance: who is responsible for defining key elements of the semantics? Can one import definitions from other communities and rely on them to maintain and document their semantics? The CSML approach has been to appropriate

[8] Full details of each CSML feature can be found in the CSML manual at http://ndg.nerc.ac.uk/csml.

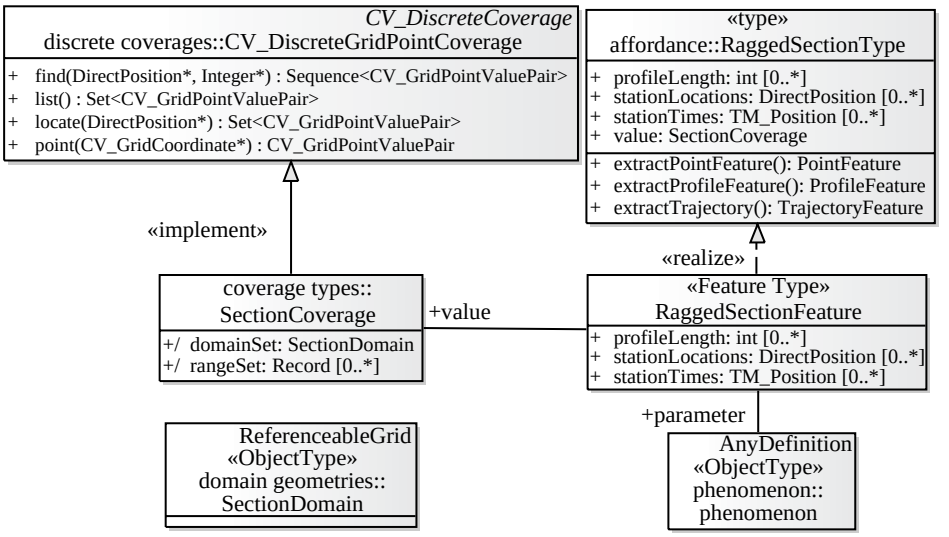

Figure 7. Example of a CSML feature type: ragged section. Note that the implementation of the methods of CV_DiscretePointCoverage is the functional attributes of the ragged section type. This pattern is replicated in all the CSML features.

phenomenon definitions from the climate forecast (CF) conventions community and from dictionaries managed by the British Oceanographic Data Centre, BODC (http://www.cfconventions.org and http://vocab.ndg.nerc.ac.uk, respectively). In both cases, management of vocabularies is dealt with by communities with well-established procedures. The vocabularies, along with ontological mappings between them, are deployed as a service (see Latham *et al.* 2009). The explicit managed ontology, along with the individual ontologies encoded within MOLES documents together form the O-metadata, which is key to traversing between discipline descriptions.

CF also introduces two important concepts for measurements of physical systems: cell bounds and cell methods. All measurements have some sort of statistical property (method) associated with them, such as whether or not they represent an instantaneous, average or maximum value. Those statistical properties are related to the coordinates, typically by the bounds representative of a measurement cell (in space and/or time). Early versions of CSML did not implement support for these concepts, but practical use of CSML has indicated their importance.

Figure 8 shows how these details of the phenomenon definition intimately link the phenomenon itself with a coverage definition. CSML is implementing support for these concepts by exploiting the Constrained Phenomenon part of the Open Geospatial Consortium (OGC) Sensor Web Enablement specifications, along with the special attributes of the coverages (CSMLCellDescription).

3. Implementation status

NDG has implemented data and metadata systems based on the above methodology; more details can be found in the companion paper (Latham *et al.* 2009).

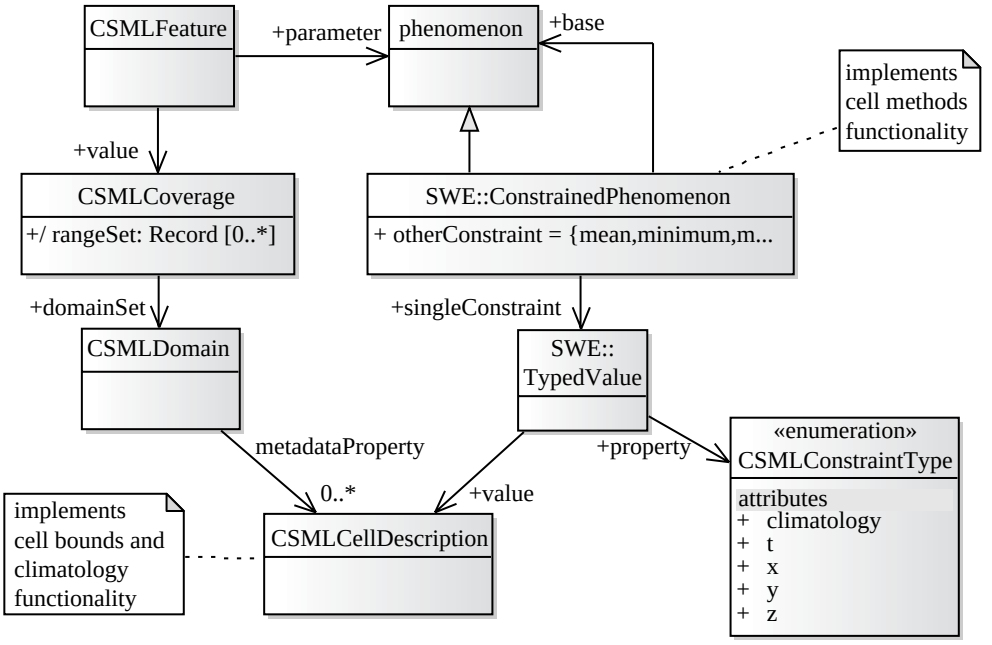

Figure 8. Implementing properties of the phenomenon inherited from the CF conventions for netCDF: describing constraints and cells.

A fully compliant NDG data provider will have data holdings and metadata, both of which are exposed via NDG services: fully compliant data providers currently include the British Atmospheric and Oceanographic Data Centres (including in the latter case both groups at the Proudman Oceanographic Laboratory and at the National Oceanography Centre in Southampton), the NERC Earth Observation Data Centre and the NERC Earth Observation Data Acquisition and Analysis Service.

NDG services deployed include reliable discovery and vocabulary services along with the experimental Web map and coverage services based on OGC protocols. While all the datasets held by the project partners (and many others) are discoverable, deployment of data access services for all datasets is predicated on populating MOLES and CSML instance documents, and so only a subset of data is currently 'on the grid'. All partners are committed to extending the number of datasets that support a full range of information and data manipulation services.

4. Further work

From an information perspective, the major thrust of activities will be to modify both the MOLES and CSML frameworks to conform to the new Observations and Measurements candidate ISO specifications (O&M; Cox 2007*a*,*b*). CSML already conforms to an extent—the design of CSML has influenced and will continue to be influenced by O&M. In addition, we will further classify the attributes and links (and hence the ontologies) available for describing the MOLES entities. NDG will be INSPIRE[9] compliant. In the particular case of

[9] The Infrastructure for Spatial Information in the European Community.

large-scale numerical simulations of the Earth system (such as climate and numerical weather prediction models), we will also work in the context of the European METAFOR project to include better classification (and hence discovery) of both model descriptions and the datasets that result from the simulations. In addition, much more needs to be done on the description of services, and the associations between services and the various information entities needed within an environmental grid.

From a deployment perspective, we will be extending the deployment of the NDG (and hence the information requirements) into the wider NERC environmental community, and beyond.

The NDG has benefited from contributions from a large team. Particularly key contributions came from Kevin O'Neill, Dominic Lowe, Kerstin Kleese Van Dam and Sue Latham. This paper has been improved by the careful reading of Simon Cox and three anonymous referees.

References

Cox, S. (ed.) 2007*a* Observations and measurements—Part 1. Observation schema. Open Geospatial Consortium technical paper OGC 07-022r1. See http://www.opengeospatial.org/standards/om.

Cox, S. (ed.) 2007*b* Observations and measurements—Part 2. Sampling features (1.0). Open Geospatial Consortium technical paper OGC 07-002r3. See http://www.opengeospatial.org/standards/om.

Foster, I. 2006 Globus Toolkit Version 4: software for service-oriented systems. In *IFIP International Conference on Network and Parallel Computing*. Lecture Notes in Computer Science, vol. 3779, pp. 2–13. Berlin, Germany: Springer Verlag.

ISO 19101 2002 Geographic information—reference model. Geneva, Switzerland: International Organisation for Standardisation.

ISO 19103 2005 Geographic information—conceptual schema language. Geneva, Switzerland. International Organisation for Standardisation.

ISO 19109 2005 Geographic information—rules for application schema. Geneva, Switzerland: International Organisation for Standardisation.

ISO 19115 2005 Geographic information—metadata. Geneva, Switzerland: International Organisation for Standardisation.

ISO 19123 2005 Geographic information—schema for coverage geometry and functions. Geneva, Switzerland: International Organisation for Standardisation.

ISO 19136 2005 Geographic information—Geography Markup Language (GML). Geneva, Switzerland: International Organisation for Standardisation.

Karasavvas, K., Antonioletti, M., Atkinson, M. P., Chue Hong, N. P., Sugden, T., Hume, A. C., Jackson, M., Krause, A. & Palansuriya, C. 2005 *Introduction to OGSA-DAI services*. Lecture Notes in Computer Science, vol. 3458, pp. 1–12. Berlin, Germany: Springer Verlag.

Latham, S. E. *et al.* 2009 The NERC DataGrid services. *Phil. Trans. R. Soc. A* **367**, 1015–1019. (doi:10.1098/rsta.2008.0238)

Wan, M., Rajasekar, A., Moore, R. & Andrew, P. 2003 A simple mass storage system for the SRB data grid. In *20th IEEE/11th NASA Goddard Conf. on Mass Storage Systems and Technologies* (*MSST2003*), *San Diego, CA, 7–10 April 2003*.